LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-572432

Finding Motifs in Wind Generation Time Series Data

C. Kamath, Y. J. Fan

August 10, 2012

International Conference on Machine Learning and Applications
Boca Raton, FL, United States
December 12, 2012 through December 15, 2012

Disclaimer

Finding Motifs in Wind Generation Time Series Data

Chandrika Kamath and Ya Ju Fan
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
Livermore, CA
Email: kamath2, fan4@llnl.gov

***Abstract*—Wind energy is scheduled on the power grid using 0-6 hour ahead forecasts generated from computer simulations or historical data. When the forecasts are inaccurate, control room operators use their expertise, as well as the actual generation from previous days, to estimate the amount of energy to schedule. However, this is a challenge, and it would be useful for the operators to have additional information they can exploit to make better informed decisions. In this paper, we use techniques from time series analysis to determine if there are motifs, or frequently occurring diurnal patterns in wind generation data. Using data from wind farms in Tehachapi Pass and mid-Columbia Basin, we describe our findings and discuss how these motifs can be used to guide scheduling decisions.**

***Keywords*-Wind Generation; Time Series Analysis, Motifs, Clustering**

I. Introduction

With renewable resources, such as wind, providing an increasing percentage of our energy requirements, integrating them into the power grid is becoming challenging. Wind is an intermittent resource; control room operators typically use 0-6 hour ahead forecasts to determine the amount of energy to schedule for the hours ahead. However, when the forecasts are inaccurate, the operators consider the actual generation in the previous hours and days, and use their experience and expertise to estimate the energy they should schedule.

In discussions on scheduling wind resources with operators at Southern California Edison, we had observed that there appeared to be a diurnal pattern in the generation for the previous days. A closer examination of historical data confirmed the presence of patterns. The generation may be low and flat on days with little wind, or it may be high in the early hours, drop down to near zero by noon, and rise again in the late evening. A similar observation was also made by operators at California Independent System Operator (CaISO), who, on days when the actual generation deviates from the forecast, use the pattern for the day thus far to find matches to patterns from previous days, which are then used to guide the amount of energy scheduled.

In this paper, we use time series analysis to identify these diurnal patterns, referred to as *motifs*, or primitive shapes [1]. Our goal is to ascertain if there is a limited number of motifs for the wind generation at a site and determine ways to exploit these motifs so operators can make better informed scheduling decisions. While many ideas have been proposed to improve scheduling, to the best of our knowledge, this is the first application of the use of machine learning to find motifs in wind generation data.

II. Description of the data

We conduct our analysis using data from two regions - the Tehachapi Pass in Southern California, which connects to the grid through Southern California Edison (SCE), and the Columbia Basin region on the Oregon-Washington border, whose wind farms form part of the Bonneville Power Administration (BPA) balancing area. We refer to these datasets as the SCE and BPA data, respectively.

A. SCE data

The SCE dataset (Figure 1) is the smaller dataset, with data for 2007-2008 sampled at 15 minute intervals for the Vincent and Antelope regions. As these regions are close by, their wind generation is very similar, and we consider the sum of the generation in our analysis. A quality check indicated a few small negative values for the generation at Antelope; these were set to zero before being combined with the value from Vincent. Note that the maximum wind generation over the two year period is constant.

Figure 2 shows the wind generation for SCE for a week in June, 2007. In this short segment of the data, there are two discernible patterns. The generation on June 1, 2, 3, and 7 starts high at midnight, drops by the middle of the day and then rises again in the afternoon. In contrast, the generation on June 4, 5, and 6, tends to remain at a consistent high level, though there is variation within each day.

B. BPA data

The BPA data (Figure 3), available for the period 2007-2011, are sampled at 5 minute intervals. There are missing values in the data - values missing for one or two consecutive intervals were filled-in using interpolation, while longer periods were replaced by "-9999" to indicate such values for future processing. Note that the maximum wind generation over the five year period has increased substantially from nearly 700 MW in 2007 to 3500 MW in 2011.

Figure 4 shows the wind generation for BPA for a week in June 2011. In this segment, we see more variation than in the segment from SCE data. June 5 starts off near zero and

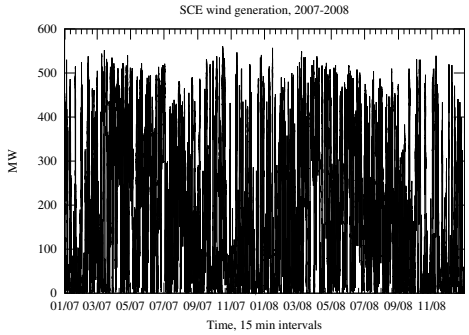

Figure 1. Wind generation in the Vincent and Antelope regions of the Tehachapi Pass, Southern California, 2007-2008.

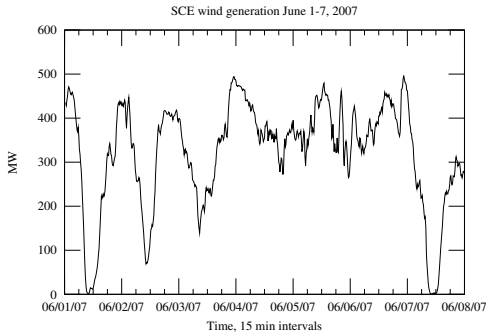

Figure 2. A week-long segment for wind generation: SCE, June 2007.

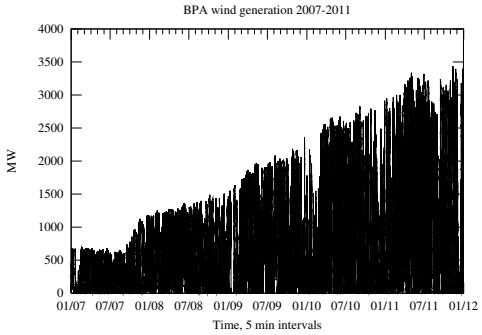

Figure 3. Wind generation in the mid-Columbia Basin region in the Oregon-Washington border, 2007-2011.

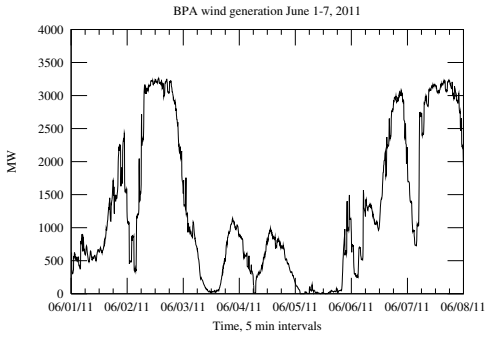

Figure 4. A week-long segment for wind generation: BPA, June 2011.

in the late evening increases to a moderate level (relative to the peak). June 1 is similar, but is shifted in the y-axis. June 2 and 7 increase in early morning to a high flat and drop by late evening. June 3 and 4 both start off at a moderate level, drop to near zero, and then increase again. However, June 4 appears to be a slightly shifted version of June 3. Finally, the wind generation on June 6 varies considerably, making it harder to identify a pattern.

III. Analysis Approach

Our goal in the analysis is to determine if there are recurring diurnal motifs in the wind generation data so control room operators can exploit them in scheduling. We consider the wind generation data as a time series, or an ordered set of m real-valued variables

$$T = t_1, \ldots, t_m \tag{1}$$

For the SCE data, $m = 70176$ (731 days) and for the BPA data, $M = 525888$ (1826 days). A subsequence S, of length n, is a subset of contiguous values

$$S = t_a, \ldots, t_{a+n-1} \tag{2}$$

from the series T. As we are interested in diurnal patterns, the value of n is chosen so a subsequence spans a day, where t_a corresponds to mid-night and t_{a+n-1} corresponds to the last time interval for which data are available for that day. Thus, for SCE, with the sampling rate of every 15 minutes, this is hour 23 minute 45 of the day, while for the BPA data sampled at every 5 minutes, this corresponds to hour 23 minute 55 of the day. For a subsequence to be considered a *motif*, it has to occur frequently enough in the time series. We will define this more precisely later in the paper.

Our analysis approach has two parts - the first represents the time series in a form more suitable for the detection of motifs and the second identifies the motifs in the data.

A. Representation of the time series

Given our time series data, it is possible to use the original data to identify the motifs; however, there are several issues with this. First, the data are noisy, and the high-frequency components in the signal can result in a large Euclidean distance between similar subsequences. Second, each day of SCE and BPA wind generation is represented by 96 and 288 values, respectively. As the data are high-dimensional, we need to map the data into a lower dimensional space, so the concept of nearest neighbors in finding similar subsequences can be meaningful [2]. Third, the increasing installed wind generation at BPA from 2007 through 2011 makes it difficult to relate a pattern in the early years to a similar pattern in the later years.

We next discuss how we address these issues, borrowing ideas from the work of Eamonn Keogh et al., in particular [1], and modifying them suitably for our datasets.

- **Scaling the data:** To account for the increasing installed capacity at BPA, we scale the 2007-2011 data by the actual installed capacity at any time, a number which is provided by BPA [3]. Once the actual generation from a new wind facility exceeds half its pending nameplate capacity, the installed capacity of the wind farms in the BPA balancing area is increased by the full nameplate capacity of the new facility. Since this is an approximation to the actual generating capacity for any day, it is possible for the maximum value of the scaled wind generation to be greater than 1.0, though not by much, as the installed capacity increases slowly. This scaling also provides us the maximum distance between any two diurnal patterns, which is the distance between a day with no wind generation and a day with the maximum wind generation for all hours in the day. We can use this maximum to set an appropriate threshold which determines when two subsequences are considered similar. For this reason, we also scale the wind generation for SCE by its maximum, even though the installed capacity does not change over the two years used in this study.
- **Piecewise aggregate representation (PAA):** Following [1], we next discretize the scaled data using piecewise aggregate approximation (PAA) [4], [5]. In this representation, each subsequence of length n (Equation 2), is transformed into a w-dimensional space by replacing n/w consecutive values by their mean. In our work, we choose $w = 24$ to correspond to the hours in a day. Thus, for the BPA data, we average $288/24 = 12$ consecutive values, while for the SCE data, we average $96/24 = 4$ values. This essentially approximates the original scaled time series with a linear combination of box basis functions [1], with the width of each box being an hour. By choosing to describe a day with 24 values, we obtain a sufficient reduction in dimensionality, without oversmoothing the data and causing loss of information during wind ramps, which are events where the generation changes by a large amount in a short time. This aggregation also reduces the noise in the original scaled data.
- **Symbolic representation (SYM):** Next, we create a discrete representation where the data are described using a few distinct symbolic values. In [1], the symbolic values are obtained by dividing the values of the time series into equiprobable bins as it allows the use of hash tables to speed up the matching process. The equiprobable bins are calculated assuming that the data follow a Gaussian distribution, which is true for most normalized time series. However, if the time series is almost constant (that is, has a fixed value corrupted by noise), the standard deviation can be quite small, and normalization by subtracting the mean and dividing by the standard deviation only amplifies the noise [6]. Our two time series have a large number of values near zero, and the resulting standardized time series do not have a Gaussian distribution. Hence, the approach in [1] is not directly applicable. Further, as we are matching one day to another, not matching all subsequences, we need fewer comparisons, and the efficiencies resulting from the use of a hash table are not essential in our problem. In light of this, we may well question the need for this additional symbolic representation. We found that this transformation can be helpful in two ways. First, it allows us to map two days with slightly differing wind generation in each of the 24 hours to the same discrete symbolic representation, enabling them to be identified as a match, while two other days which have the same difference all occurring in one hour, map to two different symbolic representations. Second, the symbolic representation allows us to include domain information in the analysis. We found that since there were many days with relatively low wind generation (less than 10% of peak for most of the day), we could capture the variations in the patterns at the low end by using more bins at the low end. This would allow us to distinguish a pattern with near zero generation for 24 hours from a pattern which started at 20% generation and slowly reduced to zero over 24 hours. Also, at the high end, we realized that the variations did not matter as much, and broader bins would suffice.
  To accommodate this, we divided our range into 10 bins, with the width of the smallest bin starting at δ and increasing by an additive factor of α each time. Thus, the bin widths for b bins are:

$$\delta,\ (1+\alpha)\delta,\ (1+2\alpha)\delta,\ \ldots,\ (1+(b-1)\alpha)\delta \quad (3)$$

  We chose $\alpha = 0.1$ and calculated δ assuming $b = 10$ bins spanning the range [0,1.0]. Any scaled wind generation values larger than 1.0 were mapped to the last bin. For ease of calculation, we represented the bins using integers, not symbols, with the distance between two bins being the difference of their integer values.

Figure 5 shows the scaled original, the PAA, and the SYM representations for the wind generation for June 1, 2011, for BPA. The horizontal lines indicate the bin boundaries for the SYM representation.

The combined PAA followed by SYM transformation is referred to as Symbolic Aggregate ApproXimation (SAX) and has been shown to be at least as good as other well known representations, such as discrete wavelet transforms and discrete Fourier transforms [7]. We considered it in our work as it is simple and provides a good representation for the patterns we see intuitively in the wind generation data.

B. Finding the motifs

Once we have the original time series converted into the reduced dimensional PAA and SAX representations, there

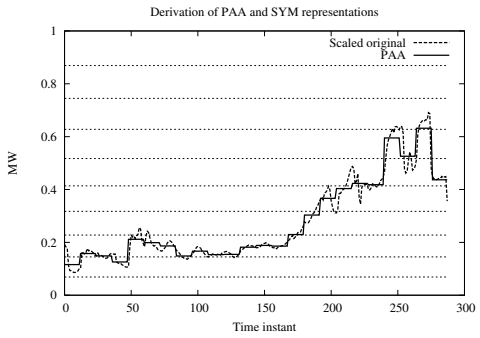

Figure 5. The wind generation for June 1, 2011, BPA, showing the scaled original and PAA representations. The dashed horizontal lines indicate the bin boundaries used for SYM. Based on this, the 24-letter SAX representation for this day would be 2-3-3-2-3-3-3-3-3-3-3-3-3-3-4-4-5-5-6-6-7-7-8-6.

are several ways in which we can find the motifs, or the frequently occurring subsequences, in the data. We considered two approaches. The first was the method proposed in [1] for finding the most frequent motif. This approach requires the setting of a range R, which is a positive real number, such that if the Euclidean distance $D(S_i, S_j)$ between two subsequences S_i and S_j is less than R, then S_j is a matching subsequence to S_i. For both PAA and SAX, we selected R to be 10% of the maximum possible distance between subsequences. Once the most frequent motif, which is the subsequence with the largest number of matches within R, was obtained, the motif and all its matches were removed from consideration and the process repeated to find the next most frequently occurring motif, and so on. We also included a constraint that for a subsequence to be considered a motif, there should be atleast 5 matches to that subsequence.

Our second approach to finding the motifs was to cluster the days in the time series. Note that the concern raised in [8] about clustering subsequences being meaningless does not apply in our case as we are interested in non-overlapping subsequences. We consider the sample-preserved k-median clustering [9], an Expectation-Maximization (EM) algorithm, in our work. This tries to find k existing samples as cluster centers such that the sum of the distances from all samples to their closest cluster center is minimized. First, a symmetric distance matrix D is calculated where the element at the i-th row and the j-th column is $D(S_i, S_j)$. Since the clusters are assigned using a distance matrix, this gives us the flexibility of choosing different distance metrics for our datasets. Then, given the number of clusters and the initial cluster centers, the algorithm alternates between an expectation (E) step and a maximization (M) step. In the E step, all samples are assigned to their nearest sample-preserved median. In the M step, the medians are reevaluated by choosing the samples that are closest to all others in the same clusters. The algorithm converges when the cluster assignments do not change across iterations.

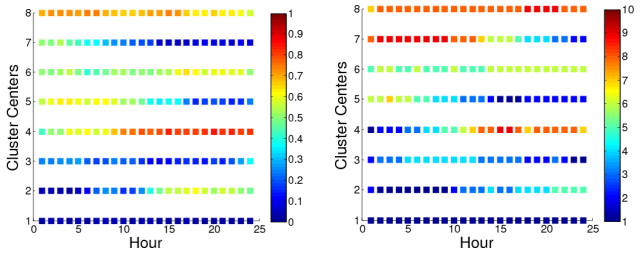

Figure 6. Cluster centers for BPA: PAA (left) and SAX (right); cluster 1 is at bottom of plot and cluster 8 is at the top.

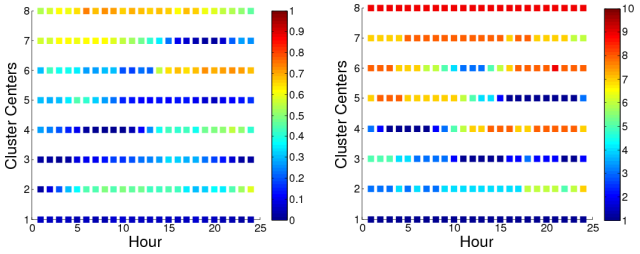

Figure 7. Cluster centers for SCE: PAA (left) and SAX (right); cluster 1 is at bottom of plot and cluster 8 is at the top.

IV. Experimental Results

We next present the results obtained using the motif finding algorithms on the PAA and SAX representations of our two time series datasets. We focus on the results from clustering, as the results using the motif finding algorithm in [1] are very similar. However, in the latter, not all days may be assigned to a motif, for example, if they are not within a distance of R of any other day, or they form a matching group of subsequences with less than 5 members.

In our work, we consider 8 clusters, as this was the number suggested by the motif finding algorithm of Lin et al. [1]. Figures 6 and 7 show the cluster centers for BPA and SCE respectively. The number of days in each cluster is summarized in Table I. The clusters for the PAA and SAX representations are shown in Figures 8 and 9, for the BPA and SCE data, respectively, in the same order as in Table I. Each cluster has the 24 hours on the x-axis, with the y-axis representing each day that belongs to the cluster. The color represents the magnitude of wind generation for that day and hour. The color palette for PAA and SAX are different as the range of values are different. A cluster represents all the days that have a similar motif.

These results show that there are indeed frequently occurring subsequences, or motifs, in the wind generation data. We also observe that the number of motifs is small and the largest cluster for both BPA and SCE, using PAA and SAX, is the one where the wind generation is quite small, as indicated by the cluster with all dark blue generation. This is also corroborated by the time series data themselves.

A closer inspection of the clusters and the number of days in each indicates several differences between BPA and SCE,

Cluster	BPA		SCE	
	PAA	SAX	PAA	SAX
8	184 (10.1%)	279 (15.3%)	173 (23.7%)	60 (8.2%)
7	170 (9.3%)	134 (7.3%)	58 (7.9%)	122 (16.7%)
6	151 (8.3%)	159 (8.7%)	72 (9.8%)	55 (7.5%)
5	170 (9.3%)	221 (12.1%)	71 (9.7%)	58 (7.9%)
4	121 (6.6%)	133 (7.3%)	61 (8.3%)	48 (6.6%)
3	274 (15.0%)	150 (8.2%)	50 (6.8%)	82 (11.2%)
2	179 (9.8%)	220 (12.0%)	45 (6.2%)	100 (13.7%)
1	577 (31.6%)	530 (29.0%)	201 (27.5%)	206 (28.2%)

Table I
CLUSTER SIZES (PERCENTAGES) FOR THE CLUSTERS IN FIGURES 8 AND 9 WHERE THE TOP CLUSTER IS CLUSTER 8.

and between the PAA and SAX representations. We expect that the motifs would be different for SCE and BPA as the two regions have very different terrain and meteorological processes. For example, unlike BPA, SCE data have days where the generation is low in the middle of the day, and moderately high at the start and end of the day. We also find that PAA and SAX tend to identify similar motifs, though given the different representations, the clusters are not identical. SAX appears to perform slightly better as it finds clusters that are clearly different, while in the case of PAA, it is unclear why some clusters (e.g., clusters 1 and 3 for both BPA and SCE) are different. We suspect that the SYM discretization in the SAX representation groups together subsequences which differ only slightly, so there is greater similarity within a cluster and the clusters appear very different from each other. For example, the cluster centers for SCE in Figure 7, tend to be more on the blue side for PAA representation, while the SAX representation is more balanced between the blue and the red.

We next analyzed the clusters to determine if there is a seasonal pattern to the motifs. For example, we found that for SCE, cluster 1 (very low generation) is more frequent in the winter (October-February), while cluster 8, with high generation for all hours, is more frequent in spring (March-May) and July. In addition to such insights, identification of the motifs also helps operators in scheduling. For example, if during the day, the operator finds that the forecast is inaccurate, they can either compare the generation with the existing motifs identified in the data, or find nearest matches in historical data. This additional information could then be used in estimating the wind power to be scheduled.

V. CONCLUSION

In this paper, we analyzed wind generation data to identify diurnal patterns, or motifs. Using time series from two wind sites, we found that there is indeed a limited number of motifs, though, as expected, these motifs may be different across sites. The motifs provide insights into the wind generation and can guide the wind energy to be scheduled when the forecast is inaccurate. We will extend this work by exploring how sensitive the results are to different discretizations in the SYM approach, the choice of clustering method, and the number of clusters. We will also investigate if removing outliers from the data will improve the clustering, and if it is possible to use weather conditions to predict the motif, an idea that appears feasible given the seasonal pattern among the motifs.

ACKNOWLEDGMENT

LLNL-CONF-572432: This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This work was part of the WindSENSE project at LLNL. We thank Jim Blatchford (CaISO), Min-Lin Cheng (SCE), Robert Farber (SCE), John Pease (BPA), Scott Winner (BPA), and John Zack (MESO) for insights into the integration of wind energy on the power grid.

REFERENCES

[1] J. Lin, E. Keogh, S. Lonardi, and P. Patel, "Finding motifs in time series," in *Proceedings of the 2nd SigKDD Workshop on Temporal Data Mining*, 2002, pp. 53–68.

[2] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *Proceedings of the 7-th International Conference on Database Theory , Lecture Notes in Computer Science*, vol. 1540. Springer-Verlag, 1999, pp. 217–235.

[3] "Bonneville Power Administration Wind Power web page," http://www.bpa.gov/corporate/windpower/.

[4] B.-K. Yi and C. Faloutsos, "Fast time sequence indexing for arbitrary lp norms," in *Proceedings of the 26th International Conference on Very Large Data Bases*, ser. VLDB '00, 2000, pp. 385–394.

[5] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," *Journal of Knowledge and Information Systems*, pp. 263–286, 2001.

[6] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2003, pp. 2–11.

[7] "Symbolic Aggregate approXimation (SAX) web page," http://www.cs.ucr.edu/~eamonn/SAX.htm.

[8] E. Keogh, J. Lin, and W. Truppel, "Clustering of time series subsequences is meaningless: Implications for previous and future research," in *Proceedings of the Third IEEE International Conference on Data Mining*, 2003, pp. 115–122.

[9] Y. J. Fan, "Optimization models and algorithms for sample-preserved classification and clustering," Ph.D. dissertation, Rutgers University, New Brunswick, New Jersey, May 2010.

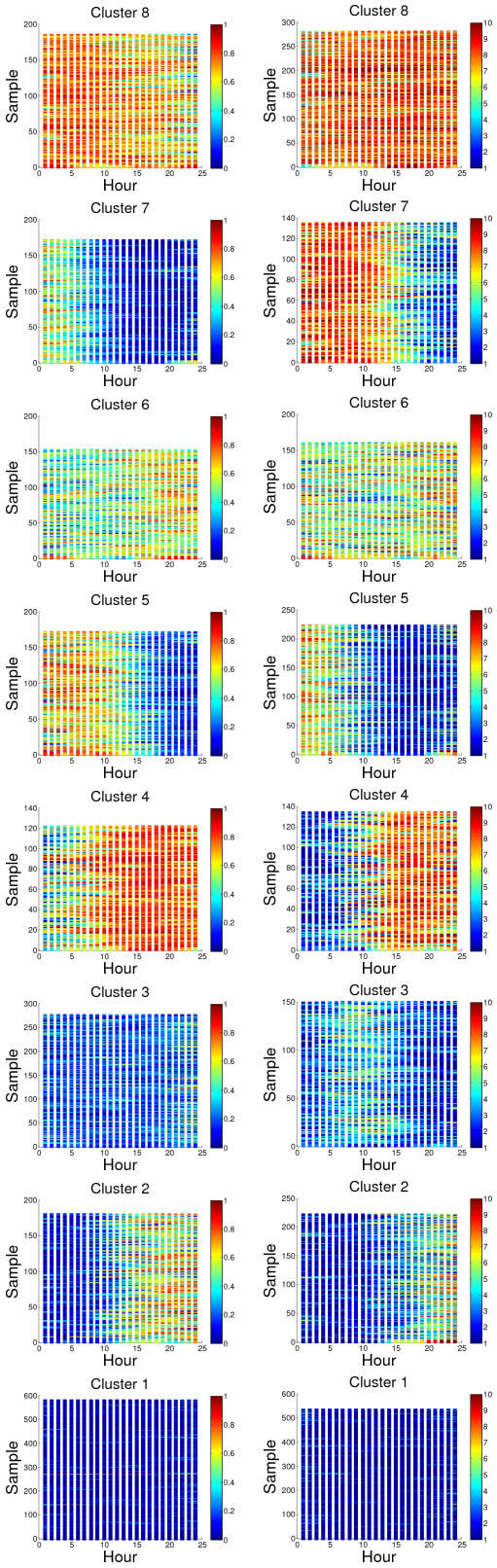

Figure 8. Clusters identified in the BPA data: PAA (left column) and SAX (right column) representations. The cluster sizes are listed in Table I. A row represents a day and the color is the magnitude of the generation. Color maps differ between PAA and SAX. Cluster 8 is at the top.

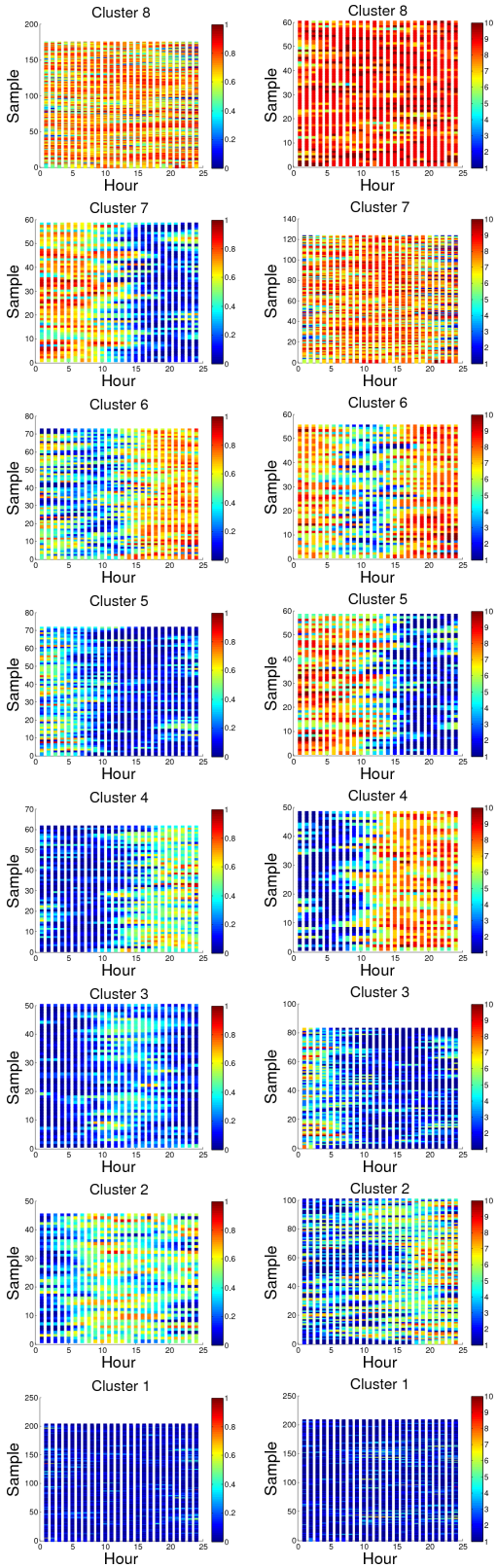

Figure 9. Clusters identified in the SCE data: PAA (left column) and SAX (right column) representations. The cluster sizes are listed in Table I. A row represents a day and the color is the magnitude of the generation. Color maps differ between PAA and SAX. Cluster 8 is at the top.